Title: Protein threading by nonlinearly combining evolutionary and non-evolutionary information

Proteins play fundamental roles in all biological processes. Akin to the complete sequencing of genomes, complete descriptions of protein structures is a fundamental step towards understanding biological life, and is also highly relevant in the development of therapeutics and drugs. Computational methods, especially template-based modeling, can quickly generate crude but useful structure models at a large scale. The challenge of template-based modeling lies in the recognition of correct templates and the generation of accurate sequence-template alignments. Evolutionary information (i.e., sequence profiles) has proved to be very powerful in detecting remote homologs, as demonstrated by the state-of-the-art profile-based method HHpred. However, there are still a lot of proteins, even in the PDB, without good sequence profiles. We present a new protein threading method for proteins without good sequence profiles by nonlinearly combining evolutionary and non-evolutionary information. In particular, we model protein threading using a probabilistic graphical model Conditional (Markov) Random Fields, which guides sequence-template alignment using a nonlinear scoring function consisting of a collection of regression trees. A regression tree estimates the log-likelihood of an alignment state from both evolutionary and non-evolutionary information. Experimental results indicate that when there is no sufficient evolutionary information, this new method greatly outperforms HHpred in terms of both alignment accuracy and mode quality, and that non-evolutionary information is helpful to around half of the templates. The paradigm presented here for the design of a nonlinear scoring function is very general. It can also be applied to protein sequence alignment and RNA alignment.